AIAA-94-1287-CP

OBJECT TRACKING WITH STEREO VISION

Eric Huber
The MITRE Corporation
Houston, TX
huber@mitre.org

Abstract

Task directed, active vision techniques are beginning to produce vision systems capable of providing real-time depth perception for robots. The concepts of shallow disparity filters and coarse disparity segmentation show particular promise for several reasons: they require minimal correlation search, thus they are fast; they employ redundant measurements, so they are noise tolerant; and they are not dependent on precise measurements, so they are calibration insensitive. We have extended these techniques by coarsely segmenting regions about the fixation point according to the task at hand. Relative occupancy of these regions provide cues to the approximate location of relevant features. They also provide reliable information for maintaining gaze stabilization. We have implemented these techniques on a real-time active stereo vision system to produce robust visual skills for tracking highly dynamic objects and occluding contours. These techniques are general in that they can be applied to a wide variety of objects.

Introduction

Vision has long been recognized as one of the most significant pieces of the autonomous robot control puzzle. The richness of information available in the visible spectrum makes vision unique from other sensing modalities [Mataric]. The vision research community initially endeavored, unsuccessfully, to assimilate all available data which resulted in computational overload. Much of the recent success in vision can be attributed to a transformation in thinking which has swept the vision community. Stereo vision, in particular, has seen a resurgence as the role of attentive mechanisms, foveation, and vergence in biological models has become better understood. Exciting developments in gaze control and task directed vision are now bringing long awaited goals within our grasp.

Classical Stereo

Stereo vision provides a means for depth determination through the principal of triangulation. If an object in a scene is located in the center of a stereo image pair, then that stereo system is said to be verged, or fixated, on the object.

In the more general case, features in a scene will be viewed from a variety of depths. Assuming that there is some mechanism for matching (correlating) features between the stereo image pair, the feature disparity (the relative number of pixels from the optical axis) for each image is required to determine the object location. In addition to the baseline, the angle subtended by each pixel must be known for the stereo cameras. The accuracy of these measurements is dependent on the accuracy with which calibrated system parameters are known.

Feature matching is a non-trivial problem in stereo vision. Extensive research has resulted in a number of schemes ranging from simple pixel-pixel correspondence or vertical line matching [Braunegg 90], to more sophisticated approaches which employ filtering of the images prior to matching in order to normalize the scene and accentuate its salient features [Marr-Poggio 79]. Still other schemes utilize a variety of primitives [Marapane-Trivedi 92] in order to determine correspondence. The approach taken to achieve correspondence is often driven by the characteristics of the domain in which the vision system is to be operated.

Stereo Methodology

When given the opportunity to acquire stereo and other forms of vision data, it is tempting to try to generate a complete 3-d map (reconstruction) of the environment. Early attempts at reconstruction consistently met with insurmountable difficulty due to the complexity of the real world (and its lighting conditions). Real-

time performance was only conceivable in extremely limited domains.

When these vision systems were applied to robotic tasks such as navigation, the lack of real-time performance became most evident [Moravec 83]. Even if an accurate reconstruction were achievable in real-time, the task of "making sense" of the resultant geometry is computationally expensive.

To complicate matters, 3-D matching algorithms require precise depth information in order to converge on a correct solution. Precise data requires precise calibration which is not only time consuming to achieve, but difficult to maintain [Grimson 93] when sitting atop a dynamic robot. Finally, biological models of successful stereo vision systems do not suggest a dependence on highly accurate "calibration".

Biological models have provided some insights to the vision community which is turning away from reconstruction and towards approaches which can be managed in real-time. The prevailing attitudes held by the vision community include:

- The desire to make computation more manageable by limiting computation to what is most relevant to achieving the task at hand [Ballard 91].

- The desire to exploit regularity in the environment to allow for reduced complexity (Successfully demonstrated) [Horswill 93].

- The need to tightly couple sensor control with what is being perceived (active vision).

The last concept is common in the stereo vision community which is particularly concerned with fixation [Ballard 91] and gaze (vergence) [Coombs-Brown 91] control.

One of the most difficult problems in stereo vision is matching; however, recent techniques for reducing the computation associated with correlation have met with success [Grimson 93]. By reducing the range over which a match is likely to occur, the search space and the false match rate are reduced [Olson 93]. This approach may be extended in order that shallow disparity regions can be used for local segmentation in depth [Grimson 93]. Such approaches have only begun to be explored but show much promise.

In support of robot activities it is necessary for vision systems to provide relevant information in real-time. Even when techniques for computational simplification have been employed, a large amount of data must be processed in order to be useful to a robot. An active, real-time stereo system is still a "complex beast". Few institutions [Nishihara 90] [Marapane-Trivedi 92] have the resources necessary to assemble and program the specialized hardware required for real time stereo vision. However, it would be more difficult to simulate the interaction between a noisy, dynamic world and an active stereo vision system than to build the real thing. These are the primary reasons for the paucity of real-time stereo vision research. This will improve as computing hardware and digital cameras become faster and more accessible.

PRISM Stereo Vision

The following discussion concerns our work in active stereo vision. The objective of our R&D efforts is to provide needed visual perception skills to the mobile and articulated robotic systems we support. This stereo vision project, which has been ongoing for approximately one year, employs a PRISM vision system which will be described briefly.

The PRISM [Nishihara 84] [Nishihara 90-B] stereo vision system is the embodiment of Keith Nishihara's biologically inspired Sign Correlation Theory [Nishihara 90-A]. Nishihara's theory is an extension of Marr and Poggio's Zero-Crossing Theory [Marr-Poggio 79] . The Zero-Crossing Theory was found to be effective at extracting the salient features of an image in a normalized fashion, but the matching algorithm did not perform well in the presence of noise. Sign Correlation is the dual to Zero-Crossing contour correlation, yet it is highly tolerant to noise and preserves the intent of Zero-Crossing Theory (Nishihara 90-B).

The PRISM system employs dedicated hardware to provide spatial and/or temporal disparities in real-time. It is comprised of two frame-synchronized cameras mounted on a servo controlled pan-tilt-verge head. The camera output is digitized and fed into a custom Laplacian of Gaussian (LOG) convolver board. The LOG convolved stereo output is then buffered on a high speed, parallel sign correlation board. A 68040 processor is used to control the operation of the LOG convolver, sign

correlator, and a motion controller for the pan-tilt-verge head.

We have selected several goals for our stereo vision system which are intended to provide the greatest benefit to the robotics projects which we support. The list is representative of the goals described by the vision community in general which includes: tracking, object size and shape estimation, obstacle detection, and object recognition. In pursuit of these goals, our approach is largely based on real-time active vision principals. In addition, we are concerned with computational economy, modularity, reusability, and man-machine interaction.

Motion-Centroid Tracking

Originally, our PRISM system included some application software for performing differential motion tracking. The tracking algorithm which was provided computes a monocular disparity vector, for a single image patch, from consecutive monocular frames (temporal disparity). The disparity vector (motion vector) is then used to update pan-tilt velocity control parameters. These parameters are fed to the motion controller which keeps the tracked object in the field of view. Vergence is maintained through stereo disparity matching and is centered about the same location as the motion correlation window.

This scheme enables the system to track an object for a surprisingly long time (due in-part to the accuracy gained by the subpixel disparity measurements made possible through sign correlation) considering that disparity errors accumulate with each frame with no mechanism for re-centering on the object. Still, these errors eventually accumulate to the point where the attentive mechanism "slips off" an occluding contour of the tracked object and "gets lost" on the background. This scheme also proves to be very fragile when faced with rotating objects.

Since the differential tracking algorithm is designed to tap the strengths of PRISM, its speed is adequate, but it lacks the robustness required for our purposes. Therefore, we developed an algorithm which combats the effects of rotation and integrated disparity error by re-centering on the object. After the motion and range disparity measurements, there was only about 1/5 of a frame period left for centering oriented measurements; thus it was necessary to stay true to our goal of computational economy.

The centering algorithm (CA) we developed is simple yet robust. It takes advantage of the fact that the differential tracker typically maintains the object of interest within a shallow stereo disparity range. The advantage to limiting disparity measurements to a shallow range is that it greatly reduces the number of correlations necessary to search a given region of space. A determination of occupancy can be made quickly by checking one of these shallow regions for high (hits) or low (misses) correlation. The centering algorithm employs a grid of shallow disparity measurements (disparity grid) which can then be interrogated to find an approximate area centroid for the object. The CA simply biases the pan-tilt velocity command towards the centroid of the object acting, in effect, as an integral control term with respect to the differential tracker.

The CA described above is almost qualitative in nature. Each measurement is binary, and a number of binary results are used to drive the attentive mechanism. This approach provides several advantages: each measurement is relatively low cost allowing for several measurements per frame; a number of measurements are averaged making the system insensitive to noise; and the measurements are well distributed, providing sufficient information to track even sparsely textured objects.

As an object is tracked, its size and shape may change as its relative pose changes, revealing different portions of the object. In order for the centroid tracking algorithm to be general, it had to be designed to normalize with respect to such image transitions. Normalization is achieved through a low cost analysis of the occupancy distribution within the disparity grid. If the "hits" and "misses" are homogeneously distributed then the grid size is increased. If there are only a few "hits" and they are localized, then the grid size is decreased . The resultant behavior is that the grid constantly seeks to match the scale of the object it is tracking, a necessary precursor to maintaining center.

A beneficial side effect of the grid size normalization mechanism is that the grid tends to expand as much as possible, thus it attempts to engulf the larger, more track-able portions of an object. For example, if a motion seeking algorithm "wakes up" the tracker when it is fixated on a person's hand, the disparity grid will tend to move the fixation point up his arm and eventually to his torso. Once fixated on the torso,

the tracker is very robust and it is difficult for someone to "lose" it.

Finally the disparity grid depth of field (defined in disparity space) had to be normalized with respect to the object's distance. Otherwise, while tracking an agent which strays far away, the expanded depth of field may allow the object to blend into the background. This also requires little more than a single trig. computation.

Depending on the task at hand, the centroid tracker can be influenced to fixate on specific regions of an object. For example, when tracking a human, a small velocity control bias in the +Tilt direction causes the tracker to reach equilibrium on the head rather than the torso. We expect this skill to come in handy in the future when we will need to segment out the head and/or face for recognition purposes.

Depth Tracking

Centroid biased motion tracking has proven effective, yet expensive. In order to find the displacement of a correlation window between consecutive frames, it is necessary to tessellate (in effect) an area of the image with correlation windows. For animate objects such as humans, anticipated accelerations require a large array of correlations. Stereo disparity measurements, however, require only a one dimensional search, thus they are relatively inexpensive.

In many instances, disparity segmentation alone is adequate for locating an object. For this purpose a depth tracking algorithm was developed which again makes use of disparity regions. This algorithm utilizes a pyramid approach to search for a distinctive object by segmenting with respect to depth and cyclopean axis proximity. Again, the object centroid is used as the reference location. For tracking spatially distinct objects, the performance of the depth tracking algorithm (implemented on the PRISM system) is astounding. In fact, in this configuration, tracking accelerations are limited by the mechanics/control of the pan-tilt-verge head rather than the computational speed. Additional visual cues would be necessary to make this approach to tracking viable in cluttered environments.

Contour Tracking

For determining the shape and size of objects which cannot be contained within the camera field of view, it is useful to be able to trace, or in a more dynamic sense, track its contours. To achieve this, the author has developed a mechanism which employs methods similar to those used for tracking object centroids. The difference is that instead of avoiding object boundaries, it is desired that the disparity grid straddle them. If the system can be made to seek out and "stabilize on contours", then determining the "sense" of the tangent vector needed to track along the contour is a trivial matter.

In order to develop a contour stabilizing algorithm using stereo disparity, it is necessary to contemplate the depth transitions characteristic of these contours. The depth gradient of a contour may be positive, zero, or negative with respect to the stereo coordinate frame.

Occluding contours have special characteristics which readily lend themselves to stable tracking. A disparity grid of measurements straddling an occluding contour will have a distinct cluster (dark area) of "hits" on an object and a distinct region (light area) of "misses" off of the object edge. An algorithm which employs competing "forces" acting to keep these areas in balance has proven stable even when limited texture was available. Contour tracking is induced by moving along the perpendicular bisector of the dark and light regions.

The remaining question is "how to update the median disparity depth of field as the contour is traversed". This question is key because the disparity depth of field must be appropriately biased to ensure that it always encompasses the occluding edge. Failure to meet this criteria will not ensure that the occluding contour will be tracked accurately as its slopes towards or away from the vision reference frame. For occluding contours, the appropriate bias is found by averaging disparities near the perpendicular bisector. The trend indicated by the depth differential between these points, and points on the interior of the edge, provide a clue about the slope of the occluding contour. The disparity depth of field must be biased in the direction of this slope.

This scheme has proven to be very successful for tracing out distinct objects (spatially

separated from other objects) including cardboard boxes (poorly textured), a robotic manipulator (wires and all), and several humans. The algorithm, as implemented on the PRISM, includes some dynamic "optimization" which modulates the head control velocity based on contour detectability and smoothness. The system can now trace out an average sized human, standing seven feet away, in about twenty seconds, although, it currently cheats in order to find closure when it reaches ground level.

Interior contours present a slightly more difficult depth segmentation problem and therefore present a greater challenge for gaze stabilization. Several approaches to real-time interior contour tracking are under development.

Conclusion

The benefits of using local disparity regions for coarse grain depth segmentation has been discussed. Several successful stereo vision skills which utilize these measurement techniques have been described.

Our PRISM system is scheduled to be integrated with a Cybermotion platform. We intend to use the visual skills described above for guiding the robot to follow humans, walking at a natural pace, through cluttered environments.

Acknowledgments

Support for research reported on in this paper was provided by NASA and was performed in the Laboratories of the Automation and Robotics Division at Johnson Space Center, NASA. I am grateful for the insight and encouragement provided by Ken Baker.

References

[Ballard 91] "Animate Vision", Dana H. Ballard, Journal of Artificial Intelligence, 48:57-86, 1991.

[Braunegg 90] "MARVEL: A System for Recognizing World Locations with Stereo Vision", David J. Braunegg, MIT A.I. Lab Technical Report 1229, May, 1990.

[Coombs-Brown] "Cooperative Gaze Holding in Binocular Vision", David J. Coombs and Christopher M. Brown, IEEE Control Systems Magazine, June, 1991.

[Grimson 93] "Why Stereo Vision is Not Always About 3D Reconstruction", W. Eric L. Grimson, MIT A. I. Memo #1435, July, 1993.

[Horswill 93] "Polly: A Vision-Based Artificial Agent", Ian Horswill, Proceedings of the Eleventh National Conference on Artificial Intelligence, 824-829, July, 1993.

[Marapane-Trivedi 92] "Multi-Primative Hierarchical(MPH) Stereo System", Suresh B. Marapane and Mohan M. Trivedi, IEEE Conference on Computer Vision and Pattern Recognition , June 1992.

[Marr-Poggio 79] "A computational theory of human stereo vision", D. Marr and T. Poggio, Proceedings of the Royal Society of London, 204:301-328, 1979.

[Mataric] "Perceptual Parallelism and Action Selection As Alternatives to Selective Perception" , Maja J Mataric (MIT Artificial Intelligence Laboratory).

[Moravec 83] "The Stanford Cart and The CMU Rover", Hans Moravec, Proceedings of IEEE, 71 :872-884, July, 1983.

[Nishihara 84] "Practical real-time imaging stereo matcher", H.Keith Nishihara, Optical Engineering, 23(5), 1984.

[Nishihara 90-A] "Real-Time Implementation of a Sign-Correlation Algorithm for Image-Matching", H. Keith Nishihara, Teleos Research Technical Report 90-2, Palo Alto, CA, 1990.

[Nishihara 90-B]"PRISM-3: Real-Time Binocular Stereo and Optical Flow Measurement System" H. Keith Nishihara, (Teleos Research, Palo Alto, CA), 1990.

[Olson 93] "Stereopsis for Vergeing Systems", Thomas J. Olson, IEEE Computer Vision and Pattern Recognition Conference, 55-60, 1993.